STATISTICAL LANGUAGE MODELING FOR SPEECH DISFLUENCIES

Andreas Stolcke *Elizabeth Shriberg*

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA 94025
stolcke@speech.sri.com
ees@speech.sri.com

ABSTRACT

Speech disfluencies (such as filled pauses, repetitions, restarts) are among the characteristics distinguishing spontaneous speech from planned or read speech. We introduce a language model that predicts disfluencies probabilistically and uses an edited, fluent context to predict following words. The model is based on a generalization of the standard N-gram language model. It uses dynamic programming to compute the probability of a word sequence, taking into account possible hidden disfluency events. We analyze the model's performance for various disfluency types on the Switchboard corpus. We find that the model reduces word perplexity in the neighborhood of disfluency events; however, overall differences are small and have no significant impact on recognition accuracy. We also note that for modeling of the most frequent type of disfluency, filled pauses, a segmentation of utterances into linguistic (rather than acoustic) units is required. Our analysis illustrates a generally useful technique for language model evaluation based on local perplexity comparisons.

1. MOTIVATION AND OVERVIEW

Speech disfluencies (DFs) are prevalent in spontaneous speech, and are among the characteristics distinguishing spontaneous speech from planned or read speech. DFs are one of many potential factors contributing to the relatively poor performance of state-of-the-art recognizers on this type of speech, e.g., as found in the Switchboard [2] corpus.

Past work on disfluent speech has focused on disfluency detection, using either acoustic features [7, 6] or recognized word sequences [1, 3]. Our goal in this work is to develop a statistical language model (LM) that can be used for speech decoding or rescoring, and that improves upon standard LMs by explicitly modeling the most frequent DF types. The main reason to expect that DF modeling can improve the LM is that standard N-gram models are based on word predictions from local contexts, which are rendered less uniform by intervening DFs. Other researchers have recently started exploring approaches to DF modeling based on similar assumptions [4, 8].

Section 2 describes a simple N-gram-style DF model, based on the intuition that DF events need to be predicted and edited from the context to improve the prediction of following words. Section 3 compares the DF model with a baseline LM, in terms of both perplexities and word error rates on Switchboard data. The emphasis is on a detailed analysis of the model at DF and following word positions. Section 4 provides a general discussion of the results.

2. THE MODEL

2.1. Disfluency types

Following [9], DFs can be classified based on how the actual utterance must be modified to obtain the intended fluent utterance, i.e., the utterance a speaker would produce if asked to repeat his or her utterance. The types can be characterized by the type of editing required.

Filled pauses (FP) The pause filler (typically "uh" or "um") must be excised.

```
SHE UH GOT REAL LUCKY THOUGH
--> SHE GOT REAL LUCKY THOUGH
```

Repetitions (REP) Contiguous repeated words must be removed.

```
IT'S A IT'S A FAIRLY LARGE COMMUNITY
--> IT'S A FAIRLY LARGE COMMUNITY
```

Deletions (DEL) Words without correspondence in the repaired word sequence must be deleted.

```
I DID YOU HAPPEN TO SEE ...
--> DID YOU HAPPEN TO SEE ...
```

We know from prior work [9] that these three types of DF are the most frequent across a variety of spontaneous speech corpora, accounting for over 85% of DF tokens in the Switchboard corpus.[1] See [9] for a description of other, less frequent, types of DF that are not modeled explicitly in our LM. For example, we are not modeling word substitutions or speech errors.

2.2. The Cleanup Model

The central assumption incorporated in our DF language model is that probability estimates for words after a DF are more accurate if conditioned on the intended fluent word sequence. A secondary assumption is that DFs themselves can be modeled as word-like events, each having a probability conditioned on the context. A standard language model, by contrast, would look only at the surface string of words and assign word probabilities in a strictly sequential manner.

Because of the central assumption, we call our DF model the 'Cleanup Model.' It is implemented as a standard backoff trigram model with the following three modifications to account for DFs.

1. Words following a DF event are conditioned on the cleaned-up, fluent version of the context. Filled pauses are removed

[1] DF frequencies in Switchboard were estimated from a hand-labeled subset of 60 conversation sides, containing 40,500 words. The coverage figure takes into account the further limits on modeled repetitions and utterance-medial deletions described below.

from contexts, as is the sequence of extraneous words in repetitions and deletions.

For example, the probability estimate for “WANT” following “BECAUSE I I” would be

$$P(\text{WANT}|\text{BECAUSE I } \textbf{REP1}) = P(\text{WANT}|\text{BECAUSE I}) \ ,$$

where **REP1** denotes a repetition event. The repeated “I” is deleted from the context.

2. Disfluencies are represented by probabilistic events occurring within the word stream, some of which are hidden from direct observation. For simplicity, we model only the most prevalent subtypes for each DF class, namely filled pauses **UH** and **UM**, repetitions of one or two words (**REP1**, **REP2**), deletions at the beginning of a sentence (**SDEL**), and other one- or two-word deletions (**DEL1**, **DEL2**).
3. Just as words, DFs are treated as events that are assigned probabilities conditioned on their context. The contexts themselves are subject to DF cleanup as described above. For example, $P(\textbf{REP1}|\text{BECAUSE I})$ is the probability of repeating “I” after “BECAUSE.”

By representing DFs simply as another type of N-gram event, we allow DFs to be conditioned on specific lexical contexts, so that simple word-based regularities in their distribution can be captured. Furthermore, because of its simple N-gram character, the model does not embody specific assumptions or constraints about the distribution of DF events.

2.3. Probability computation

To account for the hidden DF events potentially occurring between any two words, a forward computation is carried out to find the probability of a sentence prefix $P(w_1 w_2 \ldots w_k)$. Conditional word probabilities are then computed as

$$P(w_{k+1}|w_1 \ldots w_k) = \frac{P(w_1 \ldots w_{k+1})}{P(w_1 \ldots w_k)} \quad .$$

If the underlying N-gram model is a trigram, it is sufficient to keep eight states for each word position, according to whether the DF prior to w_k was **NODF** (none), **FP** (filled pause), **SDEL**, **DEL1**, **DEL2**, **REP1**, **REP2**, or the second position after a **REP2** event. To illustrate, the partial computation involving just the **NODF** and **REP1** states is shown here.

$$\begin{aligned}
P(w_1 \ldots w_k \textbf{NODF} w_{k+1}) &= P(w_1 \ldots w_{k-1} \textbf{NODF} w_k) \\
&\quad\ p(w_{k+1}|w_{k-1} w_k) \\
&\quad + P(w_1 \ldots w_{k-1} \textbf{REP1} w_k) \\
&\quad\ p(w_{k+1}|w_{k-2} w_{k-1}) \\
P(w_1 \ldots w_k \textbf{REP1} w_{k+1}) &= \delta(w_k, w_{k+1}) \\
&\quad [p(w_1 \ldots w_{k-1} \textbf{NODF} w_k) \\
&\quad\ p(\textbf{REP1}|w_{k-1} w_k) \\
&\quad + P(w_1 \ldots w_{k-1} \textbf{REP1} w_k) \\
&\quad\ p(\textbf{REP1}|w_{k-2} w_{k-1})]
\end{aligned}$$

where $\delta(w_i, w_j) = 1$ if $w_i = w_j$, and 0 otherwise. Trigram probabilities are denoted by $p(\cdot|\cdot)$; these are obtained through the usual backoff procedure [5]. The total prefix probability is then computed as

$$P(w_1 \ldots w_k) = \sum_X P(w_1 \ldots X w_k) \quad ,$$

where X ranges over the hidden states representing the disfluency types (including **NODF**).

2.4. Estimation

The backoff N-gram probabilities in the model are estimated from N-gram counts, including counts of the DF events. We used standard Good-Turing discounting in the backoff for both baseline and DF trigram models. For experiments reported here involving hidden DF events, we used a subset of the Switchboard corpus that was hand-annotated for disfluencies as well as for linguistic segments.[2] In the absence of hand-annotated training data, an iterative reestimation (EM) algorithm could be used to estimate the N-gram probabilities for hidden DF events.

When counting N-grams for the DF model, the same context modifications used in the DF cleanup operations must be performed on the training data. For example, the word sequence

```
<s> SHE UH GOT REAL LUCKY
```

is counted as having the following trigrams:

```
<s> SHE UH          <s> SHE GOT
SHE GOT REAL        GOT REAL LUCKY
```

Note that the trigrams

```
SHE UH GOT          UH GOT REAL
```

which would be generated for a standard trigram LM are *not* generated for the DF model.

Because DF and word events are represented uniformly as N-grams in the model, the standard estimation procedure will normalize DF and non-DF event probabilities. This is a convenient simplification over alternative approaches in which DFs are modeled separately from the fluent word sequences.

3. RESULTS AND ANALYSIS

3.1. Overall results

We trained a trigram model for FP, REP, and DEL disfluencies as described above, using 1.4 million words of Switchboard data labeled for DF events (see note 2). The model was then evaluated on a test set of 17,500 words. Table 1 compares baseline trigram and DF models.[3]

Table 1. Overall results

Model	Perplexity	Word error
Baseline trigram	119.1	50.21%
DF trigram	120.9	50.23%

As can be seen, there is no significant difference in recognition word error rates. While this may be due to a number of factors (some of which we discuss in Section 4), we would have expected at least a reduction in perplexity for the DF model; this was not the case. We wanted to know whether this was because our underlying assumptions were wrong, or whether it was due to other factors, so we decided to analyze the DF model performance in detail.

We note with regard to these and later results that some types of disfluencies may contain *word fragments* (from speakers cutting themselves off in mid-word). According to [9], 20 to 25% of repetitions and deletions in Switchboard contain word fragments; however, filled pauses, as classified here, never involve words fragments. Fragments are usually not part of the vocabulary of current recognizers, and are not modeled in our system. They

[2] A preliminary version of annotated Switchboard data was made available to the 1995 Johns Hopkins Language Modeling Workshop; the LDC will release a final version.

[3] Both baseline and DF models were trained on the same data, which corresponds to only a portion of the full training corpus. Therefore, the perplexity figures are higher here than in some of the comparisons below.

were therefore omitted from the transcripts used for our perplexity computations. We can expect an additional benefit from successful fragment recognition, since they would serve as extra evidence for repetitions and deletions, as well as for other DF events.

3.2. Analysis by DF type

To assess the potential of the DF model specifically at DF locations, we computed perplexities for models covering each DF type in turn, and separately for a number of word positions relative to the DF event. In each case, we also computed perplexities for the non-DF positions, to make sure the model did not penalize fluent text (reported below in the "non-DF" columns).

3.2.1. Filled pauses

A trigram model with special DF modeling for filled pauses only was trained on 1.8 million words of acoustically segmented Switchboard transcripts. The test set consisted of 1861 acoustic segments containing 17,500 words. Table 2 shows the perplexities of the baseline and FP models for the FPs themselves (UH, UM), the words after (UH+1, UM+1), and two words after (UH+2, UM+2). The surprising result is that deleting FPs from N-gram contexts does not help the LM; it actually significantly *increases* the perplexity of the word following the FP. That is, on average, the FP itself is the best predictor of the following word, not the context preceding the FP. This conclusion is also supported by the corresponding bigram perplexities, which exhibit the same pattern. Apparently, FPs correlate strongly with certain lexical choices or syntactic structures, and thus give useful information regarding their neighbors to the right. We investigate this question further in Section 3.3.

3.2.2. Repetitions

A trigram model with DF modeling for repetitions was trained and tested as described above. Table 3 shows word perplexities for positions relative to repetition events. REP1 refers to the second instance of a repeated word in a one-word repetition; REP1+1 and REP1+2 denote the first and second word, respectively, after such a repetition. REP2 and REP2+1 refer to the repeated words in a two-word repetition; REP2+2 denotes the word following a two-word repetition. Unlike the case of FPs, the Cleanup Model is generally beneficial in REP contexts, reducing the joint perplexity (all the above positions relative to REP) from 85.9 to 76.6.

We also tested whether the words following the repetition might be better predicted by the REP event itself, rather than the actual words being repeated, analogous to what we found for filled pauses; this turned out not to be the case.

3.2.3. Deletions

A deletion-only DF model was trained on 1.4 million words of DF-annotated transcripts. In order to perform an analysis by DF type and position, the models were tested on 17,800 words of similarly annotated data.[4] Only sentence and one-word deletions are reported in Table 4, since the test data contained only a single two-word deletion. The second row for "DEL model" gives the perplexity based only on the word following a deletion, without including the probability for the deletion event itself. This shows that the context modification has the intended effect of making the next word more likely on average.

3.3. Filled pauses and utterance segmentation

As shown above, the Cleanup Model as applied to filled pauses yields a higher perplexity overall than the baseline trigram model. This is largely attributable to poorer word probability estimates at locations immediately following a filled pause. In prior work Shriberg [9] observed that filled pauses tend to occur at linguistic segment (e.g., clause) boundaries. Since the standard LM test utterances are segmented according to acoustic criteria, filled pauses around linguistic boundaries can actually occur in the middle of acoustic utterance segments. At such locations, the assumptions of the Cleanup Model would be grossly violated, since the preceding words actually belong to a different linguistic segment. The standard model, on the other hand, can produce reasonable predictions, as the filled pause can serve as an indicator of the boundary.

To test this hypothesis we compared the perplexities of both models on a subset of the test data that was hand-annotated for linguistic segmentations, and that had been re-segmented accordingly (10250 words in 1325 segments). Specifically, we compared the perplexities of words following *medial* filled pauses, i.e., filled pauses not occurring as the first or last word in a linguistic segment. Results are shown in Table 5.

Table 5. Local perplexities after medial filled pauses

Position	UH+1	UM+1
Baseline	849.0	437.4
FP model	606.2	361.7

We see that the Cleanup Model is the better predictor for words following medial FPs, the reverse of the result for acoustically segmented utterances. That is, the cleanup assumption holds for medial FPs if one models utterances based on linguistic, rather than acoustic, segments.

3.4. Results from related work

We are aware of two other groups of researchers currently investigating similar approaches to DF language modeling. In [4] a cleanup-style model for filled pauses is described. Ries and Qui at CMU [8] have experimented with models for repetitions and certain types of sentence deletion that incorporate the cleanup assumption. Overall, their results are consistent with ours (higher perplexity for filled pauses, lower perplexity for repetitions and deletions), but the overall effects are small, as in our case.

4. DISCUSSION AND CONCLUSIONS

The preceding analysis shows that a disfluency model based on the intuition underlying the Cleanup Model can yield only very small improvements in model perplexity, although the cleanup assumption seems to be valid on the Switchboard data we used in our experiments. The local perplexity analysis we performed shows that the word positions at and immediately following DF events can be predicted with sometimes significantly lower perplexity, although the effect on overall perplexity is very small, due to the low frequency of DF events.

An interesting (and *prima facie* unexpected) result was that the Cleanup Model does not lower perplexity for filled pauses in acoustically segmented utterances. We attribute this to the particular way that the cleanup assumption is violated by filled pauses at linguistic segment boundaries internal to an acoustic segment. There are correlations between segmentation and other types of DFs, too, but the effects on the LM should be smaller in those cases as the bigram contexts for following words are not as radically changed by different segmentations. Our findings highlight the need for a more careful modeling (possibly with automatic recovery) of linguistic structures in conversational speech, a topic we plan to address in future work.

However, even for repetitions and deletions, it does not follow that recognition accuracy would necessarily improve with better local perplexities. In fact, we tested a trigram DF model (modeling only REP and DEL events) against a standard trigram on a Switchboard test set of 1192 segments, and found virtually no difference

[4] Due to differences in amount of training data and type of segmentation, the perplexities are not directly comparable to the previous two studies.

Table 2. Local perplexities at filled pause positions.

Position	UH	UH+1	UH+2	UM	UM+1	UM+2	non-FP	overall
Baseline	39.0	223.5	89.8	174.9	36.7	71.9	103.4	101.9
FP model	39.9	291.5	91.4	175.8	73.4	69.2	103.4	103.3
#events	502	502	373	188	188	94		19426

Table 3. Local perplexities at repetition DF positions.

Position	REP1	REP1+1	REP1+2	REP2	REP2+1	REP2+2	non-REP	overall
Baseline	47.7	183.4	86.4	111.0	12.6	216.0	102.9	101.9
REP model	38.2	191.5	84.7	94.2	2.1	222.3	103.1	101.3
#events	386	386	320	44	44	44		19426

Table 4. Local perplexities at deletion DF positions.

Position	SDEL	SDEL+1	DEL1	DEL1+1	non-DEL	overall
Baseline	415.5	49.5	523.3	35.0	75.5	76.2
DEL model	402.0	47.9	544.2	36.0	75.4	76.1
word event only	99.1		289.5			
#events	130	130	15			20454

in overall word error rate (49.5% in both cases). This can be attributed to a number of factors. First, the REP/DEL model affects only a small portion of the total corpus (less than two cases per 100 words). Second, its advantage in modeling REP/DEL contexts should rarely come into effect due to the high error rate on adjacent words.

There are other reasons why lower perplexity may not lead to reduced word error rate. For instance, it could be that DFs tend to involve words of high frequency for which good acoustic models exist, so that a slightly improved LM would not affect recognition accuracy.

The overall conclusion is that by DF modeling at the LM level, contrary to high hopes in parts of the LM community, one should not expect a significant improvement in terms of word recognition performance. The main reason is that DFs are inherently local phenomena that are modeled surprisingly well by standard N-grams, even without context "cleanup."

On the positive side, our results confirm that DFs have a systematic, nonrandom distribution that can be partly captured even with simple N-gram-like models; it is therefore conceivable that more sophisticated approaches could reap benefit for recognition accuracy.

One potential source of improved DF modeling are correlations with speaker identity. For example, [9] found that speakers can be grouped into those preferring deletions over repetitions ('deleters'), and those with the opposite tendency ('repeaters'). Such cross-utterance effects could be modeled in the LM using standard techniques, e.g., using adaptive interpolation of specialized models.

Finally, we note that the language modeling techniques described could also be used for automatic disfluency tagging and removal. Given a sequence of words and a probabilistic DF model of the type used here, one can use a Viterbi-style backtrace to recover the most likely sequence of DF events underlying the words sequence. This is another application we plan to study in the future.

ACKNOWLEDGMENTS

We would like to thank Roni Rosenfeld and his LM96 group for making crucial resources available. Thanks are also due Mari Ostendorf and Patti Price for helpful discussions and comments. This research was supported in part by the Advanced Research Projects Agency under NSF Grant IRI-9314967. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the National Science Foundation.

REFERENCES

[1] J. Bear, J. Dowding, and E. Shriberg. Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 56–63, University of Delaware, Newark, Delaware, June/July 1992.

[2] J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings IEEE Conference on Acoustics, Speech and Signal Processing*, volume I, pages 517–520, San Francisco, March 1992.

[3] P. Heeman and J. Allen. Detecting and correcting speech repairs. In *Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics*, pages 295–302, New Mexico State University, Las Cruces, NM, June 1994.

[4] R. Isotani and S. Matsunaga. A study of handling filled pauses in statistical language modeling. In *Proc. Acoustical Society of Japan*, pages 81–82, 1994. [In Japanese].

[5] S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3):400–401, March 1987.

[6] C. H. Nakatani and J. Hirschberg. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America*, 95(3):1603–1616, 1994.

[7] D. O'Shaughnessy. Correcting complex false starts in spontaneous speech. In *Proceedings IEEE Conference on Acoustics, Speech and Signal Processing*, volume I, pages 349–352, Adelaide, Australia, 1994.

[8] K. Ries, 1995. Personal communication.

[9] E. E. Shriberg. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, Department of Psychology, University of California, Berkeley, CA, 1994.